Diffusion Models for Adversarial Purification

Year: 2022 | Citations: 571 | Authors: Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat

---

Abstract

Adversarial purification refers to a class of defense methods that remove adversarial perturbations using a generative model. These methods do not make assumptions on the form of attack and the classification model, and thus can defend pre-existing classifiers against unseen threats. However, their performance currently falls behind adversarial training methods. In this work, we propose DiffPure that uses diffusion models for adversarial purification: Given an adversarial example, we first diffuse it with a small amount of noise following a forward diffusion process, and then recover the clean image through a reverse generative process. To evaluate our method against strong adaptive attacks in an efficient and scalable way, we propose to use the adjoint method to compute full gradients of the reverse generative process. Extensive experiments on three image datasets including CIFAR-10, ImageNet and CelebA-HQ with three classifier architectures including ResNet, WideResNet and ViT demonstrate that our method achieves the state-of-the-art results, outperforming current adversarial training and adversarial purification methods, often by a large margin. Project page: https://diffpure.github.io.